SPEECH INTELLIGIBILITY IN COMMUNICATIVE DIFFICULTY

Sallyanne Palethorpe

Speech, Hearing and Language Research Centre
Macquarie University

An inability to maintain adequate intelligibility in situations of communicative difficulty may be due to a speaker being unaware of the requirements of the listener for maintaining intelligible conversation. In the context of an interaction with a simulated automatic speech recognition system, providing directed feedback from the listener as to the source of the communicative failure was an effective method of overcoming the problem.

INTRODUCTION

In most everyday situations, speakers can communicate with their audience quite comfortably by adapting their speech to the prevailing environmental conditions and type of audience. For example, they will talk more loudly and more carefully to overcome increased background noise or modify their speech when talking to children as opposed to adults. Lindblom (1987) has argued that this "strategy of adaptive variability" is evidence for the goal-oriented or teleological nature of the speech production process: that is, speakers adapt their speech production strategies to the demands of on-line speaker-listener interaction. Indeed, previous studies have shown that speakers are able to promote the intelligibility of their speech in order to overcome certain communicative difficulties (Chen, 1980; Picheny *et al.*, 1985; van Summers *et al.*, 1988).

However, this is not always the case. There are several well-attested examples of communicative situations where speakers utilise speech production strategies which are quite inappropriate for promoting intelligibility; for example, shouting (Rostolland, 1985). Both Lehiste (1973) and Cutler (1990) have suggested that the inability of some speakers to produce intelligible speech in certain communicative contexts may be due to the fact that they are unaware of the specific requirements of the listener in that particular setting. As most of these inappropriate production strategies are seen in situations of communicative difficulty or failure, it may well be that in these conditions, additional information is needed by some speakers to maintain intelligible communication.

Some support for this hypothesis comes from work with the hearing- and language-impaired communities, where appropriate feedback from the listener as to the cause of such communicative failure has been shown to assist speech intelligibility (Till and Toye, 1988). This feedback usually takes either a general form such as 'What?' or a more directed form which gives some indication of where the communicative failure might lie, thus allowing the speaker to revise or reformulate the utterance appropriately.

Pisoni *et al.* (1984) and van Summers *et al.* (1988) have suggested that directed feedback could also be provided from an automatic speech recognition (ASR) system to speakers about their articulation and how it could be selectively modified to improve speech recognition performance. There has been little research in this area, although it is generally recognised that man-machine communicative often takes place under stressful and difficult communicative conditions where recognition is not optimal. The aim of the present research was to see whether speakers could be trained, using feedback, to adapt their production strategies successfully to promote intelligibility whilst in communication with a simulated ASR system.

METHOD

The research methodology consisted of two experiments: the first was a production experiment which generated tokens that were subsequently tested in an intelligibility experiment. In the production experiment, the speakers had to read words to what they believed was an ASR system. They received feedback, apparently in response to each utterance but in fact predetermined by the experimenter, which either indicated that the word had been correctly recognised, or asked them to repeat it because it was not recognised. The 'repeat' condition was subdivided according to whether no indication was

given of what was 'wrong' with the utterance (the general feedback or whole word [WW] condition), or whether the computer claimed to have 'heard' another word differing by just one sound, for example lark instead of luck, thus focussing the speaker's attention on a particular phonetic contrast (the directed feedback or minimal pair [MP] condition). In addition, the test words were read in a list by the speakers and recorded for the Control condition. Five speakers were involved in the task and the test materials were six pairs of words differing by one phoneme placed in various positions within the word: luck/lark, bat/bet, gaze/daze, tab/dab, rice/rise and lease/leash.

In the intelligibility experiment to be reported here, five tokens were selected at random from the ten recorded for each test word in the control condition, the initial reading to the simulated ASR system (the initial response [IR] condition), and the two types of feedback conditions, WW and MP. Because the intelligibility testing involved only twelve test words repeated several times, some under conditions which are thought to encourage clarity of speech, it was necessary to mask the tokens using speech-shaped noise in order to provide a sensitive listening test. With a large number of tokens to be tested, it was impractical to examine them at a variety of signal to noise (S/N) ratios. Therefore a preliminary intelligibility study was carried out to determine an adequate S/N ratio for each of the individual word types for subsequent presentation in the main intelligibility study. It was not thought essential to have the same word types heard at the same S/N ratio, as the principal area of comparison was across conditions, not between words. Foils were also recorded by other speakers for use in the intelligibility tests .

As shown in Figure 1, the intelligibility testing was divided into two due to the large number of tokens involved.

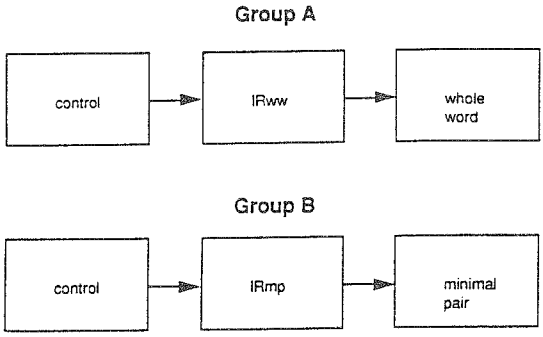

Figure 1. Protocol for the intelligibility experiment

The distinction between the two groups was based on whether the initial response to the simulated ASR system was followed by the general feedback (whole word or WW) or directed feedback (minimal pair or MP). The Control condition used the same tokens for each group, but the tokens in the other two conditions were different for Group A and B. Thus each group contained five tokens of each of the 12 test words x five speakers x three conditions, giving 900 test words plus 300 foils. Two sets of 12 listeners heard either Group A or B, with the tokens randomised across conditions and presented over three sessions in six different orders of presentation. The form of the test was open-response with the listeners writing down what they heard. The results were expressed as percentage intelligibility (intelligibility scores) and words were scored correct only if all phonemes matched the intended word.

RESULTS

As the word types varied considerably, certain restrictions were placed on the presentation of the results. There is no lack of information demonstrating the effect of the linguistic properties of the test

materials on intelligibility scores (Kalikow *et al.*, 1977). Therefore, it was not unexpected that the individual word types would have different baseline mean intelligibilities, and as a consequence, the results were examined for each word type. In addition, the intelligibility scores were averaged across speakers as, although there were the expected differences in percentage intelligibility between individual speakers, the percentage changes that occurred between conditions were similar. Detailed statistical analyses of the intelligibility scores for the Control and IR conditions demonstrated that the listeners in each group could be presumed to be sampled from the same larger population of listeners and, as a consequence, the results from both the Control and IR conditions could be pooled across Groups A and B, where necessary, and those from the WW and MP conditions directly compared.

Figure 2 shows the increase in intelligibility from the Control condition to the IR condition where the speakers believe they are communicating with an ASR system. While the baseline figure varied between words, the increase averaged eighteen percent across all words. T-tests showed that the differences were significant, with p values less than 0.001 for all words except Rice ($t=3.09, p<0.01$), Rise ($t=2.39, p<0.05$) and Bat ($t=2.59, p<0.05$). As the intelligibility test was an open-response task, it was not appropriate to look at phonemic confusions; however, a brief look at the errors showed that most of the improvement in the intelligibility was associated with an increase in intelligibility of the initial consonant.

Figure 3 displays the difference in intelligibility scores between the two types of feedback: general (WW condition) and directed (MP condition). Statistical tests produced t values that were all greater than 2.02, to give p values less than 0.05. On average, this increase was approximately 16 percent, although it did vary considerably between words. The lower intelligibility for the WW condition was in fact not different from the intelligibility for the Control condition, while the intelligibility scores for the MP condition were similar to, but not better than, those of the IR condition.

Chi-square analyses of the error scores were used to test whether the increase in intelligibility for the MP condition was brought about by an increase in the intelligibility of the phoneme contrasted. Table 1 shows that there was indeed a significant increase in intelligibility for ten of the possible twelve phonemes contrasted.

Word	Initial Consonant	Vowel	Final Consonant
Daze	(*)		
Gaze	(-)		
Tab	(*)		
Dab	(**)		
Lease			(*)
Leash			(*)
Rice			(**)
Rise			(***)
Bat		(*)	
Bet		(-)	
Lark		(**)	
Luck		(**)	
Significance	*** p <.001		
	** P <.01		
	* p <.05		
	- nonsignificant		

Table 1. Comparison of proportion of correct phonemes between WW and MP conditions

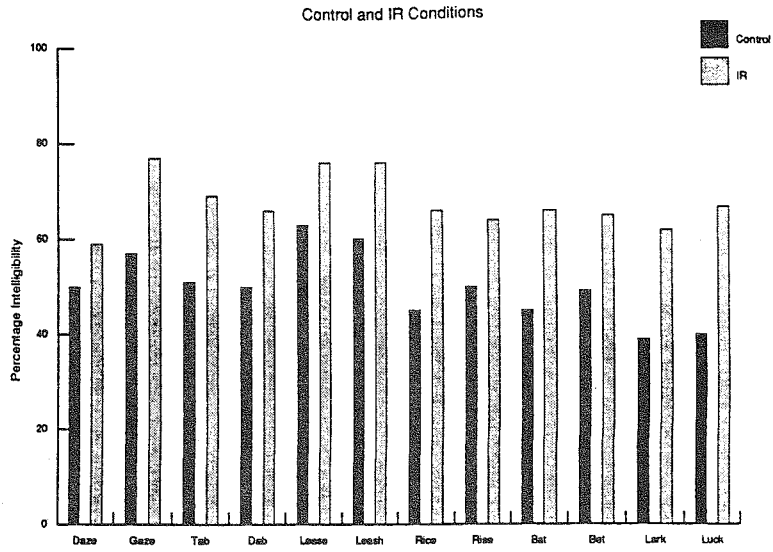

Figure 1. Comparison of mean intelligibility scores between Control and IR conditions

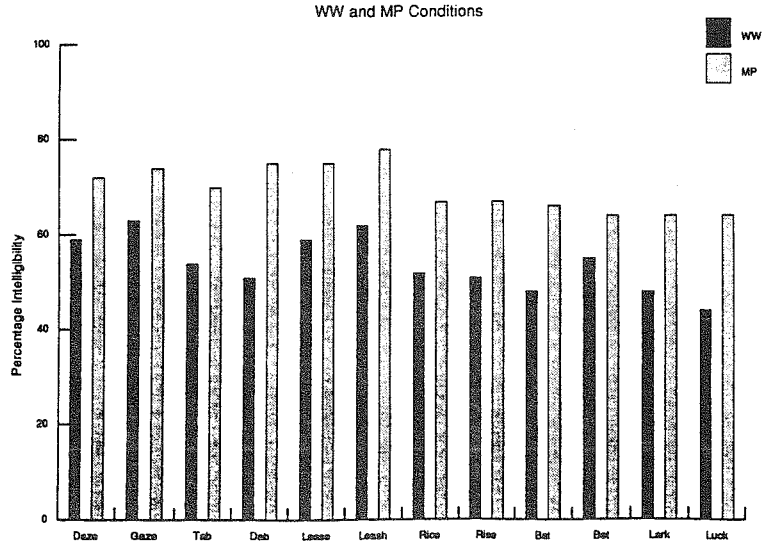

Figure 2. Comparison of mean intelligibility scores between WW and MP conditions

Further examination of the statistical results showed that those differences were in the direction of an increased intelligibility for the contrasted phoneme in the MP condition. Those word types that showed no significant differences were Gaze and Bet. One possible explanation is that these two test words had the smallest differences in intelligibility scores between the WW and MP conditions, approximately ten percent; as this difference was only just significant, it may be that the threshold was too low to allow for differences to show up at the phoneme level in the error scores.

CONCLUSIONS

The results show that when speakers believe themselves to be in communicative difficulty, as when interacting with a simulated ASR system, they are able to effectively promote the intelligibility of their speech. However, speakers are liable to become less intelligible when asked to repeat a word where recognition has failed; if you are able to provide sufficient feedback to enable speakers to perhaps locate the source of their error, then speech intelligibility improves, due mainly, in this case, to an increased intelligibility of the contrasted phoneme itself. These findings have some practical importance in suggesting a need to incorporate more directed feedback into ASR systems. On the theoretical side, these results offer only limited support for a teleological view of speech production where a speaker is expected to tailor the clarity and explicitness of his speech to the listener's needs; perhaps the speaker may be less able to achieve this by instinct and more dependent on well-directed feedback than is generally assumed.

REFERENCES

Chen F.R. (1980) *Acoustic characteristics and intelligibility of clear and conversational speech at the segmental level.* Unpublished M.S. Thesis, Massachusetts Institute of Technology, Cambridge, Mass.

Cutler A. (1990) *From performance to phonology: Comments on Beckman and Edward's paper.* In Papers in Laboratory Phonology 1: Between the Grammar and Physics of Speech, Kingston J. and Beckman M.E. (eds.). Cambridge: Cambridge University Press, 208-214.

Epstein A., Giolas T.G. & Owens E. (1968) *Familiarity and intelligibility of monosyllable word lists.* Journal of Speech and Hearing Research, 11, 428-434.

Lehiste I. (1973) *Phonetic disambiguation of syntactic ambiguity.* Glossa, 7, 107-122.

Lindblom B. (1987) *Adaptive variability and absolute constancy in speech signals: two themes in the quest for phonetic invariance.* Perilus, V, 2-20. Institute of Linguistics, University of Stockholm.

Picheny M.A., Durlach N.I. & Braida L.D. (1985) *Speaking clearly for the hard of hearing I. Intelligibility differences between clear and conversational speech.* Journal of Speech and Hearing Research, 28, 96-103.

Pisoni D.B., Bernacki R.H., Kubaska C.A. & Nusbaum H.C. (1984) *Talking in noise: Acoustic correlates of increased vocal effort and implications for speech recognition.* Research on Speech Perception, Progress Report No., 10, 169-196. Department of Psychology, Indiana University.

Rostolland D. (1985) *Intelligibility of shouted voice.* Acustica, 57, 103-121.

Till J.A. & Toye A.R. (1988) *Acoustic phonetic effects of two types of verbal feedback in dysarthric subjects.* Journal of Speech and Hearing Disorders, 53, 449-458.

Van Summers W., Pisoni D.B., Bernacki R.H., Pedlow R.I. & Stokes M.A. (1988) *Effects of noise on speech production: Acoustic and perceptual analyses.* Journal of the Acoustical Society of America, 84, 917-928.

A NEW APPROACH TO SPEAKER ADAPTATION BY MODELLING THE PRONUNCIATION IN AUTOMATIC SPEECH RECOGNITION

Florian Schiel

Lehrstuhl für Datenverarbeitung,
Technische Universität München, Germany

ABSTRACT – To deal with large lexica (more than 2000) many systems of automatic speech recognition (ASR) use an internal phonetic representation of the speech signal and phonetic models of pronunciation from the lexicon to search for the spoken word chain or sentence. Therefore there is the possibility to model different pronunciations of a word in the lexicon. In German language we observed that individual speakers pronounce words in a typical way that depends on several factors as: sex, age, place of living, place of birth, etc. Our goal is to enhance speech recognition by automatically adapting the models of pronunciation in the lexicon to the unknown speaker. The obvious problem is: You can't wait until the present speaker will have uttered approx. 2000 different words at least one time. We solved this problem by generalization of observed rules of differing pronunciation to not observed words.

Another point presented is speaker adaptation by re-estimating the a-posteriori probabilities of the phonetic units used in a 'bottom up' ASR system. A word hypothesis is evaluated by the product of the a-posteriori probabilities of the phonetic units produced by the classification to the phonetic units belonging to the word hypothesis. Normally these probabilities are estimated during the training of the ASR system and stay fixed during the test. We propose a algorithm which observes the typical confusions of phonetic units of the unknown speaker and adapt the a-priori probabilities. The learning rates can be dynamically adjusted by the entropy of the a-posteriori probabilities. By that we achieve a very fast adaptation of the a-posteriori probabilities to the optimal recognition rates using a Maximum-Likelihood criterion.

1. INTRODUCTION

Most systems of automatic speech recognition (ASR) achieve very good results even for very large vocabularies, if they are trained and used in a speaker dependent mode. But the recognition results decrease considerably if they are used in speaker independent mode. To close the gap between speaker dependent and speaker independent mode many ASR systems use algorithms to adapt the system to the unknown speaker. Most of these algorithms found in literature try to transform the preprocessed speech signal or to adapt the models of speech used in the classification algorithm. All these algorithms are able to adapt to the typical production of phonetic units and smaller events, including varying dynamics in time and loudness, pitch etc. of the unknown speaker.

Another point is the typical pronunciation of words by the unknown speaker. Most systems of ASR use a phonetical transcription of the words in the lexicon, either to combine phonetic units according to that transcription for recognition ('top down') or to evaluate word hypothesis by comparing the results of the classification with the transcription ('bottom up'). In most cases these transcriptions are drawn automatically from the orthographic representation of the words, eg. by a lookup in a lexicon of pronunciation or by algorithms of speech synthesis. Therefore in most systems of ASR there exists a unique transcription of each word of the lexicon.

Of course only very few speakers (typically well trained wireless announcers) speak a certain word in the way these ASR systems use to represent and therefore expect it to be spoken. In German language we observed the following:

- Some speakers use to pronounce certain phonemes in a special context in a different way, eg. the word 'Berlin' is correctly spoken as (phonetic symbols according to [SAM, 1990]):